The SRI-ICSI Spring 2009 Meeting Recognition System

Andreas Stolcke Kofi Boakye Adam Janin

Dimitra Vergyri Gokhan Tur

SRI International, Menlo Park, CA, USA

International Computer Science Institute, Berkeley, CA, USA

Funding: DARPA/CALO (SRI), AMIDA & IM2 (ICSI)

Thanks: Gerald Friedland, David Imseng, Haizhou Li, Trung Hieu, Marijn Huijbregts, Jon Fiscus, Jerome Ajot

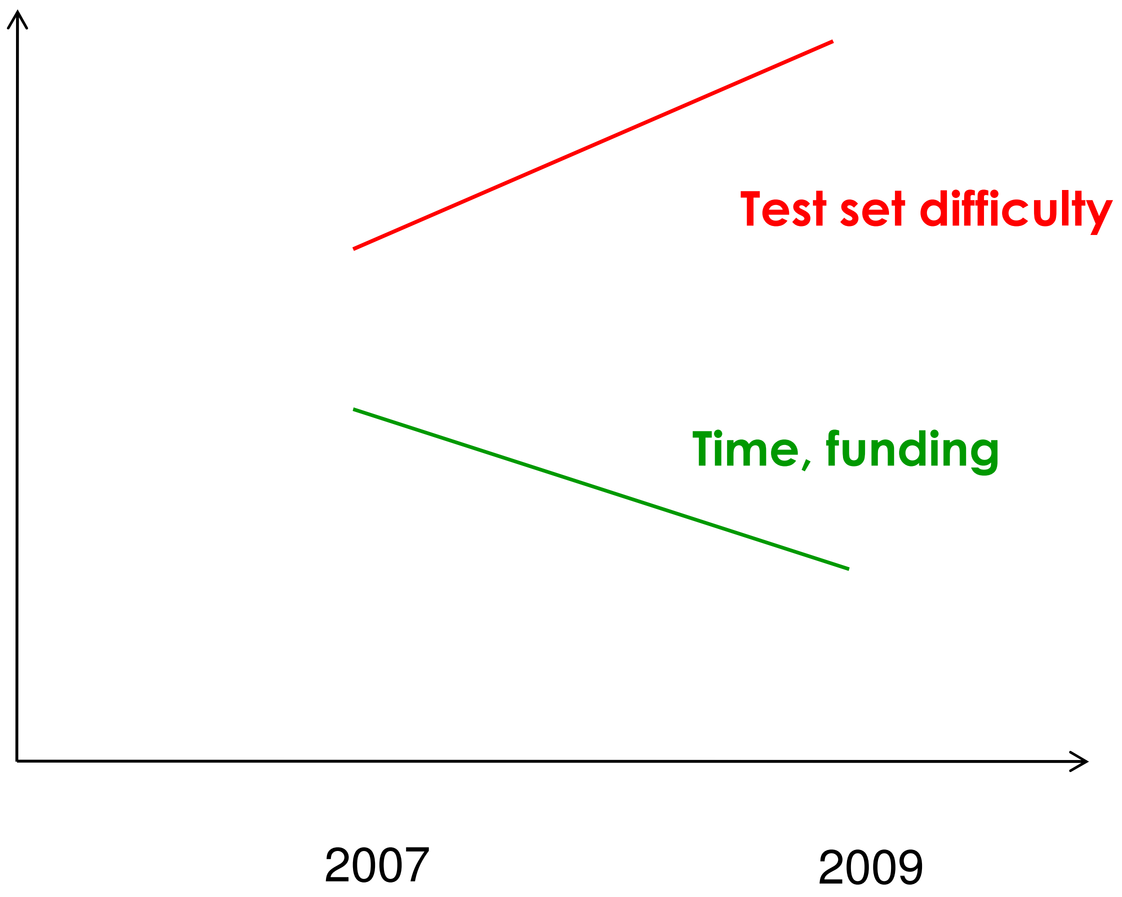
Overview

- What's new ?
- System overview
  - Architecture
  - Acoustic preprocessing
  - Acoustic and language models
- Improvements in IHM recognition
- Improvements in distant mic recognition
- Speaker-attributed recognition
- CALO-MA: meeting recognition in the wild
  - Live recognition
  - Partially supervised LM adaptation
- Summary and conclusions

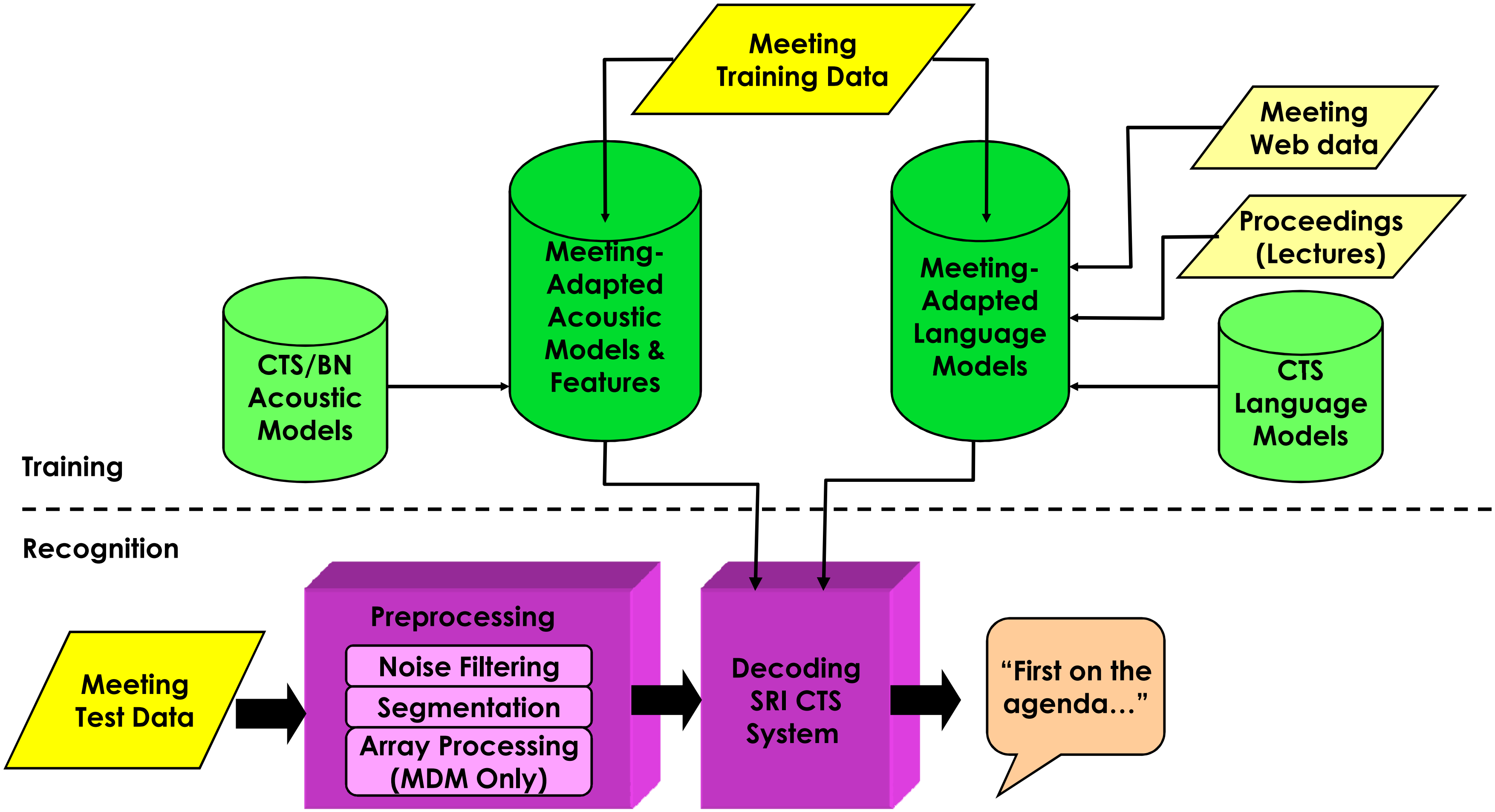
What's New?

- Very limited effort for RT-09 (2 person-weeks)
  - No new training data processed
  - Focus on better segmentation and speaker clustering
  - Heavy use of system combination (CPU cores are so cheap now …)
- Some acoustic modeling work for IHM
  - Utilized alternative acoustic model set in system combination
  - Tried to incorporate bandwidth mapping (Karafiat '08) – but failed
- Same SDM/MDM models as in RT-07
- Use of diarization for SDM/MDM
- First-time official SASTT submission
  - Error model
- CALO real-time, live ASR system
  - Not evaluated in RT-09

This Year's Challenge



Meeting STT System



Acoustic Preprocessing

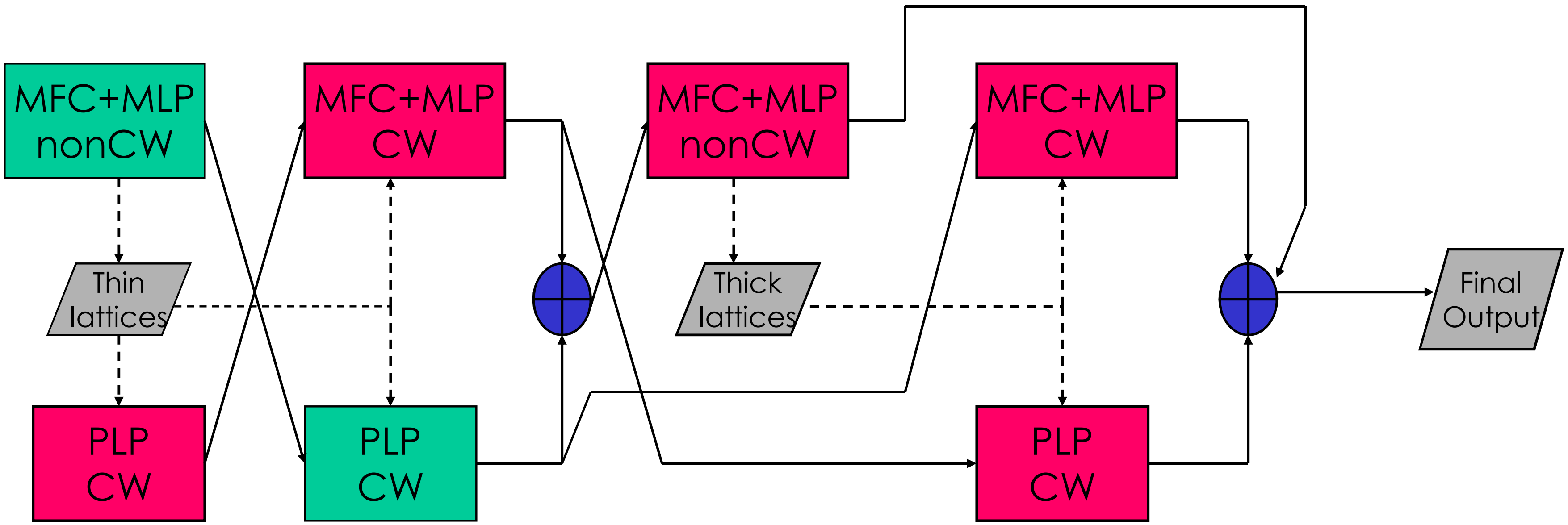
- IHM
  - HMM speech/nonspeech segmentation and cross-talk suppression with augmented cross-channel energy features (Boakye & Stolcke, 2006)
- MDM, MM3A
  1. Per-channel noise reduction with ICSI-Qualcomm Aurora Wiener filter
  2. Delay-sum processing with Xavi Anguera's BeamformIt 2.0 (same as in '07)
  3. HMM segmentation
  4. Bottom-up pseudo-speaker clustering based on GMM mixture weights

  OR

  3.' Speech/nonspeech from ICSI diarization system (plus merging/padding)

  4.' Speaker clusters from ICSI diarization system
- SDM
  - Same as MDM, without beamforming

Basic Decoding Architecture



Legend

Decoding/rescoring step

Hyps for MLLR or output

Lattice generation/use

Lattice or 1-best output

Conf. Network combination

Runtime versus Accuracy

- No Gaussian shortlists, no speed tuning in eval system
- Runtimes taken on Intel 3.0 GHz, 2x4-core CPUs
- Results for RT-09 IHM data:

System	Decoding passes	WER	Runtime
One-stage (includes segmentation)	1	32.1	0.9 xRT
Two-stage	2	28.0	1.2 xRT
Multi-stage (see diagram)	8	27.3	3.3 xRT
2 x multi-stage, combined	16	25.5	6.4 xRT

- Runtime for RT09 MDM data: 7.5 xRT

Meeting Datasets

- Development: **eval06, eval07** (confmtg data only)
- Testing: **eval09**
- Meeting training data (same as for RT-07)
  - AMI (170 meetings, 100 hours)
  - CMU (17 meetings, 11 hours) – Lapel personal mics, no distant mics
  - ICSI (73 meetings, 74 hours)
  - NIST (27 meetings, 28 hours) – **did not process newly released data**
- Acoustic background training data (same as for RT-07)
  - CTS (Switchboard + Fisher, 2300 hours)
  - BN (Hub-4 + TDT2 + TDT4, 900 hours)

Acoustic Models (from RT-07)

- Two sets of models chosen for complementary strengths, effective system combination
- MFCC + MLP models
  - Telephone front end (8kHz sampling)
  - Adapted from CTS baseline models
  - Gender dependent
  - ICSI phone-posterior features appended, estimated by multi-layer perceptron
- PLP models
  - Full-band front end (16kHz sampling)
  - Adapted from Broadcast baseline models
  - Gender independent
- Training procedure
  - ML-MAP estimation on meeting data, from MPE background models
  - fMPE-MAP feature transform estimation (Zheng & Stolcke, 2007)
  - MPE-MAP adaptation

Language Models (from RT-07)

- Linearly interpolated mixture N-gram LMs
  - Different N-gram orders for different decoding stages
  - Perplexity optimized on held-out data (AMI, CMU, ICSI, NIST)
  - Final LMs entropy-pruned
  - Vocabulary: 54k words

- Conference meeting LM components
  - Switchboard + Fisher CTS (30M words)
  - Hub4 and TDT4 BN transcripts (140M)
  - AMI, CMU, ICSI, and NIST meeting transcripts (2M)
  - Web data selected to match Fisher (530M) and meeting (382M) transcripts

Updated IHM Segmentation

- Raised prior probability for speech detection
- Augmented cross-channel energy features
  - Old: **min** and **max** of differences in normalized log energies b/w channels
  - New: added **mean** and **range** of log energy differences
- Revised training data and model configuration
  - Using all 2007 training data (added AMI training data); realigned references
  - Increased number of Gaussians per model (from 512 to 2048)
- Results using 2-pass IHM system, eval07 data:

Segmentation	WER
Baseline (RT07 auto segmentation)	30.4
+ retuned speech prior (post-RT07)	28.9
+ augmented x-channel features	28.4
+ revised training configuration	28.1
Reference segmentation	27.3

Expanded IHM Model Combination

- Old IHM acoustic models
  - MFCC: CTS-based, fMPE-MAP feature transforms + MPE-MAP training
  - PLP: BN-based, fMPE-MAP feature transforms + MPE-MAP training
- Alternate acoustic models (trained for CALO system)
  - No fMPE transforms, only MPE training (for speed)
  - PLP models CTS-based (because of limited bandwidth)
  - Non-native CTS speakers used in base models (instead of in MAP adaptation)

	eval06	eval07	eval09	eval09-refseg
Old models	20.1	23.3	27.3	24.1
New models	20.1	23.7	29.0	25.3
System combination	19.4	22.8	**25.5**	**23.8**

Submitted results

- Model combination very effective on auto segmentation
  - 1.8% absolute gain over old models (only 0.3% on reference segments)
  - 1.7% absolute gap between reference and auto segments
  - IHM test data is getting progressively harder each year …

Diarization for STT

- In past years, we were never able to get a gain from using diarization in STT preprocessing
- Our "**standard**" approach:
  - HMM speech/nonspeech segmentation
  - Bottom-up clustering into 4 pseudo-speakers per meeting
- Found in post-RT07 work: gains from combining subsystem based on *different speaker clusterings*
- Cambridge U.: broadcast recognition benefits from combining *alternate segmentations*
- **New** approach:
  - Make diarization segmentations and clustering work for STT
  - Combine with standard approach

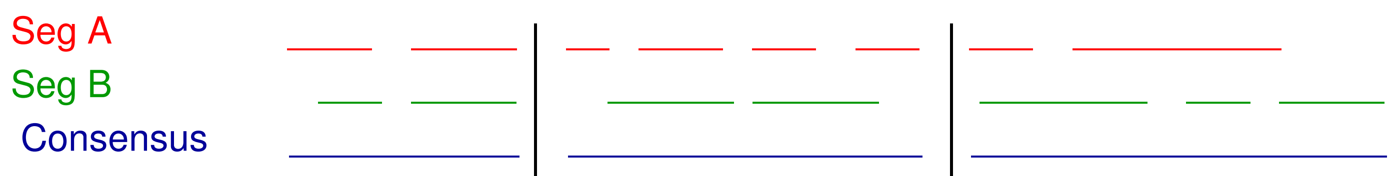
Diarization for STT (continued)

- Developed based on ICSI SPKR system
- Speech segments are merged, padded, and filtered
  - Parameters tuned on eval06 MDM
  - Merge segments by same speaker, separated by less than 0.4s nonspeech
  - Add 0.2s nonspeech around each segment
  - Remaining segments shorter than 0.2 s are discarded
- Diarization speaker clusters used for VTLN, cepstral normalization, and MLLR
- Results with overlap=1

	eval06 MDM	eval06 SDM	eval07 MDM	eval07 SDM
Standard (seg + cluster)	30.3	40.6	26.2	33.1
Diarization seg + std clustering	29.5	40.8	26.4	33.1
Diarization (seg + clustering)	29.3	39.3	25.9	32.5

Combining Multiple Segmentations and Speaker Clusterings

- Combine standard and diarization-based systems
- Baseline approach: NIST ROVER on 1-best outputs
  - Voting based on word confidences
  - Works even though input systems use different segmentations
- Better approach: Confusion network combination
  - Resegment hypotheses at gaps agreed upon by both systems



  - Concatenate, then combine confusion networks according to consensus segs

System	eval07 MDM	eval07 SDM	eval09 MDM	eval09 SDM
Standard	26.2/40.5	33.1/45.2	34.0/42.9	41.3/49.9
ROVER-combination	25.0/37.5	31.9/43.8	34.2/43.8	42.2/51.1
CN combination	24.9/37.4	31.3/43.6	33.3/43.0	40.8/50.1

Results for overlap=1/overlap=4

Primary submission

Contrast (late) submission

Effect of Diarization Quality

- Diarization-based STT worked well on eval07, but was a loss on eval09
- STT seems to degrade as a function of diarization error
- CN combination with standard system fairly robust
- Tried additional diarization systems (thanks!) with STT
  - Segment smoothing parameters were NOT retuned

Segmentation / clustering	eval09 MDM		eval09 SDM	
	DER	WER	DER	WER
Standard	n/a	34.0/42.9	n/a	41.3/49.9
ICSI diarization	17.2	35.9/43.9	31.3	44.6/51.6
IIR/NTU diarization	9.2	34.7/43.4	16.0	40.9/49.4
Standard + IIR/NTU	n/a	**32.7/41.5**	n/a	**40.0/48.8**

WERs for overlap=1/overlap=4

A Shot at Overlapping Speech

- If diarization could detect overlapping speakers …
- STT could potentially recognize overlapping speech aided by
  - Speaker-specific LM contexts
  - Acoustic models adapted to speakers' non-overlapping speech
- Quick experiment with AMI diarization system that explicitly labels overlapping speakers

Segmentation / clustering	eval09 SDM
Standard	41.3/49.9
AMI diarization w/o overlap	41.9/50.2
AMI diarization with overlap	41.9/50.2

WERs for overlap=1/overlap=4

- With MDM, we could explore beamforming speaker-specific delay estimates

MM3A Results

- MM3A data processed the same as MDM
- No special tuning performed
- Blind beamforming on all channels

Segmentation / clustering	Signal	eval09 WER
Standard	Delay-sum	43.0/56.2
Based on ICSI diarization (DER = 28.3)	Delay-sum	42.8/55.2
ROVER combination	Delay-sum	42.1/54.9
Standard	Single mic	39.4/53.9

WERs for overlap=1/overlap=4

Primary submission

- Surprise: diarization helped in spite of high DER
- Surprise: single array mic better than delay-sum
  - Need to sanity-check beamformer

Speaker-attributed STT

- Script merges STT CTM and SPKR RTTM output by assigning speaker label to each recognized word
  - Chose longest overlapping speaker if speaker change falls within a word
  - If word falls outside speech region detected by diarization, assign most recent speaker label
  - Developed by Chuck Wooters post-RT07

	Diarization system	eval09 MDM	eval09 SDM
ROVER combination	ICSI	38.2/47.7	53.6/60.9
CN combination	ICSI	37.7/47.3	52.7/60.3
CN combination	IIR-NTU	33.6/42.8	43.3/53.1

Primary submissions

Contrast (late) submissions

SASTT errors for overlap=1/overlap=3

SASTT Error Model

- Do SASTT errors behave as expected?
- Assuming SPKR and STT errors are independent, we can predict SASTT word error rate as

$$WER_{SASTT} = WER_{STT} + CorR_{STT} \times (ME_{SPKR}+SE_{SPKR})$$

where

WER_{STT}	is word error rate
$CorR_{STT}$	is word correct rate
ME_{SPKR}	is speech miss error rate
SE_{SPKR}	is speaker labeling error rate

SASTT Error Model Results

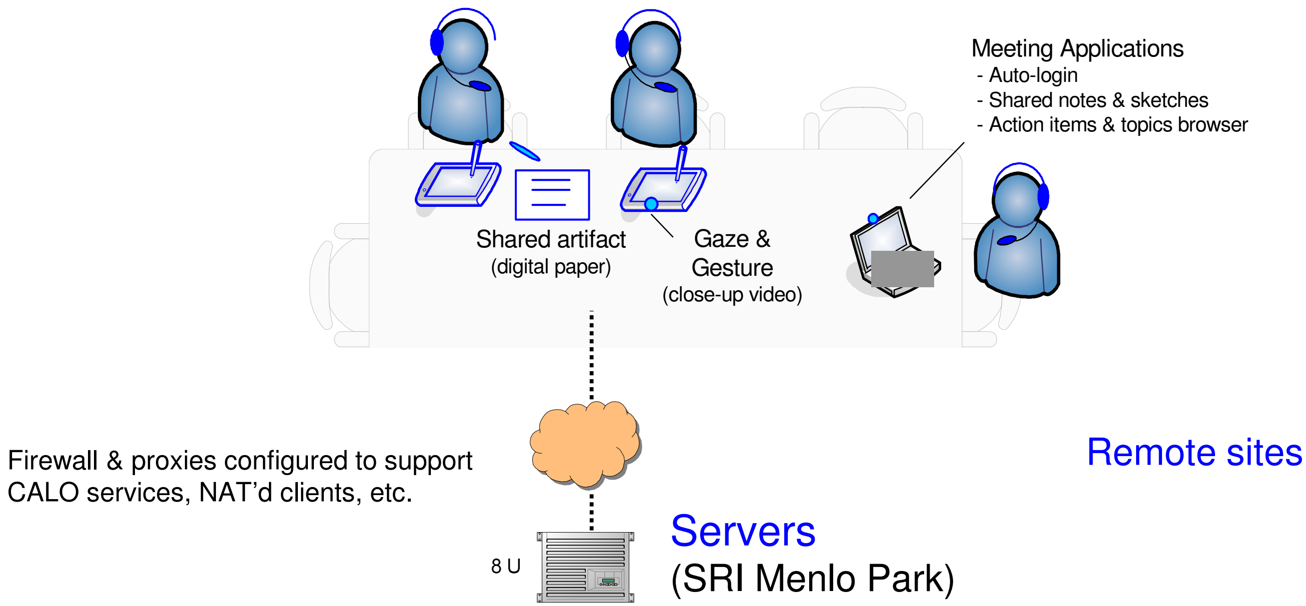
- eval09 system, IIR/NTU diarization, overlap = 1

Error metric	MDM	SDM
STT WER / WCorR	32.7 / 70.0	40.0 / 62.4
Diarization ER / ME / SE	3.8 / 0.7 / 1.2	10.7 / 0.7 / 8.2
SASTT WER predicted	34.0	45.6
SASTT WER actual	33.6	43.3

- Prediction works very well for MDM, okay for SDM (found similar results on RT07 outputs)
- SASTT error is over-estimated
- Suggests that STT and SPKR errors are correlated (conditions leading to poor ASR also cause problems for diarization)

CALO Meeting Assistant

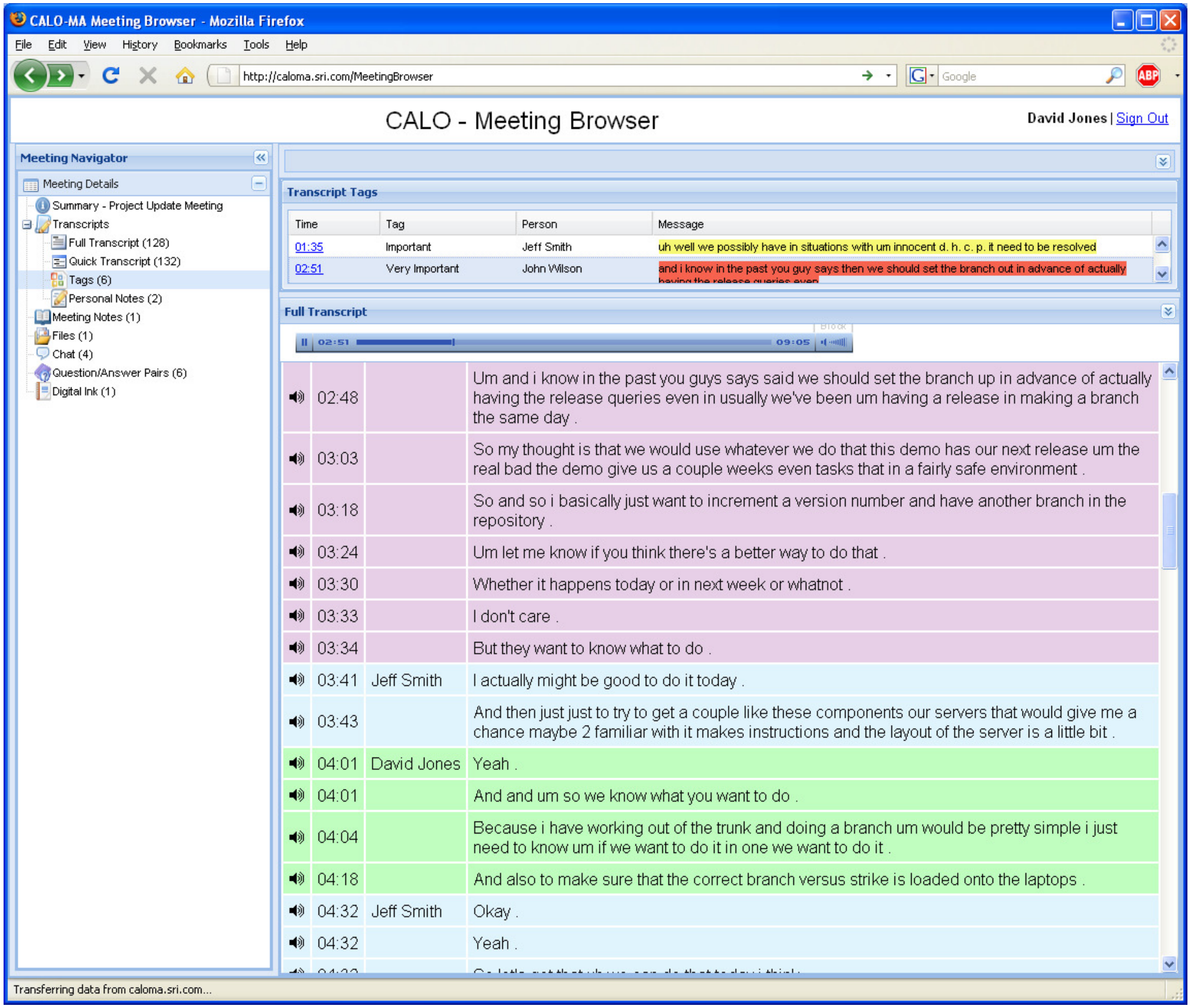
Meeting Recognition and Understanding



- Both *live* and *batch mode* recognition systems
  - Live output used for providing information to the user on the fly
  - Batch mode used for providing rich transcript of previous meetings
- System is *adaptive*: improves with use

CALO Meeting Browser

Meeting review interface



CALO-MA Recognition

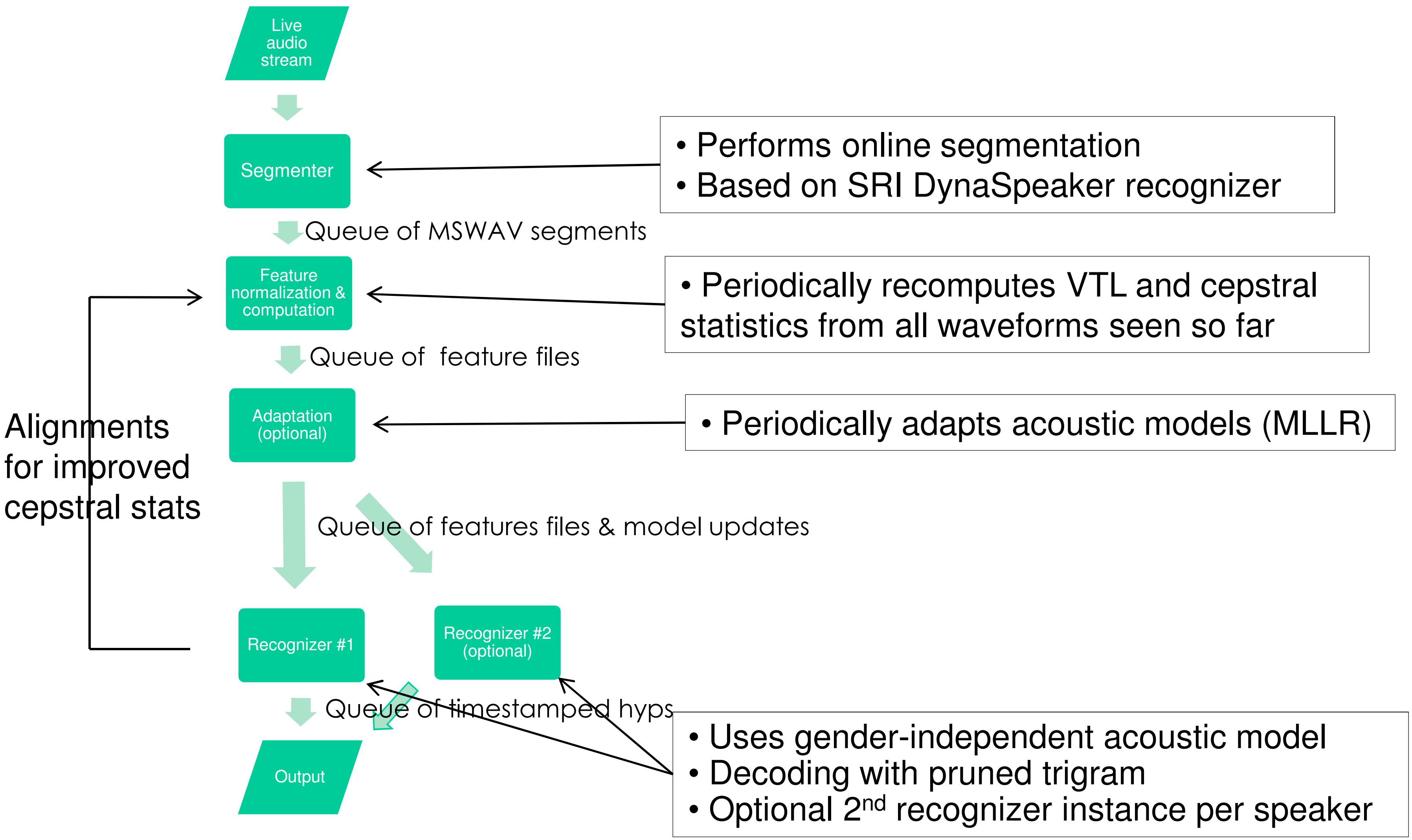
Batch recognizer

- RT-07 decoding structure (minus 1 decoding pass that gives little gain)
- CTS-based acoustic models (to deal with bandwidth limitation)
- Gaussian shortlists
- Runtime: 1.7xRT on 3GHz, 2x4-core CPU

Live recognizer

- Recognizes utterances as soon as they are endpointed
- Causal VTLN and cepstral normalization
- Causal MLLR (background process updates acoustic models periodically)
- Gender-indep. PLP acoustic models
- Pruned trigram LM
- 1-pass decoding
- Run-time: ~ 1xRT on 1CPU core
- Latency: ~ 5 - 15 seconds

CALO-MA Live STT Architecture



CALO-MA Recognition Performance

- Live recognition accuracy suffers from three factors
  - Simpler models and algorithms
  - On-line cepstral normalization
  - On-line segmentation
- Results on Sept. 2006 CALO-MA IHM data
  - Difficulty comparable to NIST eval sets

	WER
Batch system	26.0
1-pass batch system	32.5
Live system w/batch segmentation	39.6
Live system w/live segmentation	40.9
+ online adaptation	39.7

Exploiting User Feedback: (Semi)-Supervised LM Adaptation

- Principal idea:
  - Give the user the option to make corrections to ASR output from previous meetings
  - System can learn from user feedback:

    Use the (partially) corrected output to adapt the LM used for follow up meeting sequences
- Why would users provide corrections?
  - Users typically look at the output in order to remember details/prepare for following meeting
  - May be motivated to make corrections to *improve readability*
  - More motivated to *fix errors in transcription of their own* speech (which is seen by other users)
  - Partial corrections are more probable, typically covering *important/content words*
- For details see Vergyri et al., ICASSP '09

Simulated User Feedback: Partial Corrections

- Assume users correct most frequent/important content word errors.
- Assume users DO NOT correct spurious function (or stop) word errors UNLESS they are part of a larger sequence of errors.
  – E.g.: "*joined kayla project*" (errorful region)
  "*join the calo project*" (corrected)
  – Function word *the* is also restored

- Simulate various levels of correction effort
  – Randomly choose error regions to correct
  – Vary percentage of errors fixed

Experimental Setup

- Collected 8 sequences of meetings
  - Each sequence contains up to 5 meetings
  - Total of 35 meetings: ~32K words
  - Each sequence contains meetings on the same topic (e.g., hiring new staff)
  - 10 speakers in total, re-occurring across meetings
- Evaluated system improvements using across-sequence LM adaptation
  - Train LM on sequences 1-4
  - Tune weights on sequences 5-6
  - Test on sequences 7-8
  - Compared different adaptation methods, for unsupervised semi-supervised and fully supervised adaptation

Results with Varying Degrees of Feedback

- All results with linearly interpolated adapted model
- WER looks at all word errors. For semantic processing (IR, MT, summarization),content words are more important.
- The goal of user feedback is to fix as many content words as possible: look at content-WER (cWER) vs function-WER (fWER)

% total words corrected	% cont. words corrected	% WER	%cWER (rel. improv.)	%fWER (rel. improv)
0 (no-adapt)	0 (no-adapt)	16.1	12.0	19.4
0 (unsup)	0 (unsup)	15.4	11.3 (6%)	18.5 (4.6%)
15	25	15.0	10.8 (10%)	18.3 (5.7%)
30	50	14.7	10.4 (13%)	18.0 (7.2%)
55	100	14.0	9.4 (22%)	17.6 (9.3%)
100 (sup)	100 (sup)	14.0	9.4	17.5 (10.3%)

Summary & Conclusions (1)

- IHM: significant gains from improved segmentation
- IHM: modest gains from additional acoustic models and expanded system combination
- MDM/SDM: now using diarization system segmentation and speaker labels
  - Gains on eval07, and eval09 as long as diarization is sufficiently accurate
  - System combination with standard system give additional gains
  - Especially with hypothesis resegmentation and confusion network combination
- Significant improvements in all conditions masked by increasing difficulty of test data (last 3 evaluations)
- SASTT error model
  - Predicts SASTT error well based on diarization and STT error stats
- Future work: recognize overlapped speech
  - Need diarization that labels overlapping speech!
  - Run cheating experiment using reference diarization

Summary & Conclusions (2)

- CALO Meeting Assistant
  - Real-time live recognition and
  - Batch-mode post-meeting recognition
  - Semantic recognition: dialog acts, question/answer pairs, action items
- Partially supervised adaptation based on user feedback
  - Correcting about 50% of the errors (all content word errors) we achieve the same result as with fully supervised adaptation.
  - By correcting on 30% of the errors (focusing on content words) we achieve half the maximum improvement
- Future work
  - Incremental, unsupervised or partially supervised acoustic adaptation
  - Unsupervised LM adaptation with web data
  - Evaluate live recognizer using NIST evaluation data and framework

Thank You!

SRI International

ICSI
INTERNATIONAL COMPUTER SCIENCE INSTITUTE